NAVAL POSTGRADUATE SCHOOL

Monterey, California

DELAYS AT A FACILITY WITH DEMAND

FROM MANY DISTINCT SOURCES

by

Donald P. Gaver

August 1973

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral M. B. Freeman
Superintendent

M. B. Clauser
Provost

ABSTRACT

In this paper approximate models are given to describe backlogs of demands at a service system confronted by finitely many users. An application is to a time-shared computer center. Various functionals of the number waiting are discussed using the approximation.

DELAYS AT A FACILITY WITH DEMAND FROM MANY DISTINCT SOURCES

D. P. Gaver
Naval Postgraduate School
Monterey, California

1. Introduction.

Many computation centers and other job-shop-like systems, e.g. repair shops, are characterized by time-dependent random demand, and by sometimes considerable work backlogs. Computation centers that service terminals also have the peculiarity that if all customers are currently waiting in queue for attention, or are receiving service in a round-robin fashion, then no further demands can take place. This feature is characteristic of many "repairman" type problems, the latter being prototypes of numerous cyclic queueing situations; see Feller [4] for a classical treatment, and Gordon and Newell [6] for other developments.

In this paper several new approximate models are given for the computer center problem. The models explicitly confront the time-dependent demand problem, where the latter refers to daily demand if desired. The results are easy to understand and compute with; they apply most accurately when the number of system users is large, as will often be true.

2. Assumptions.

Let m terminals be connected to a main computer facility. At time t suppose that the rate of new demand from each terminal is $\lambda(t)$. Effectively we are assuming that during hour t the total number of demands is a pure birth process (see Feller [4]), or approximately Poisson for short time intervals (non-time-homogeneous), provided the terminals are occupied and their users are not awaiting computational return (service) from the facility.

Assume that the rate of completion of active user interactions is a constant, μ, so that if a number of users are awaiting or undergoing computational attention at time t, the chance that one leaves is μdt to terms of order dt. This assumption is in keeping with a first-come, first-served discipline; more importantly, it conforms approximately to a round-robin discipline, in which each customer present receives a small quantum of computer time.

3. Model Equations.

The above assumptions imply that $Q(t)$, the number of terminal users that await or undergo service at time t, is described by a birth and death stochastic process, see Feller [4]. One can envision--but not relish--the task of writing down $m + 1$ simultaneous differential equations for the probabilities that $Q(t) = j$ $(j = 0,1,\ldots,m)$, working out a solution, and then interpreting the results.

The solution must necessarily be numerical, and will be tedious and expensive to obtain if m grows large. We present here an approach that is approximate but should be quite accurate when m is moderately sizable, and business is brisk at the computer. The approach is based on approximating $Q(t)$ by a deterministic part, plus an additional noise, the latter being characterized as a random process. In the present situation the noise turns out to be a familiar Gaussian diffusion process, the Ornstein-Uhlenbeck, see Cox and Miller [1]. It has properties that may be exploited in order to shed light on the computer's service capabilities.

If m, the number of terminals, is large then it is plausible that $\frac{Q(t)}{m} \to q(t)$, a deterministic function. That the latter is true for constant $\lambda(t)$, equal to λ, follows from results of Iglehart, for the closely related repairman problems; see [7]. Next consider

$$N_m(t) = \frac{Q(t) - mq(t)}{\sqrt{m}} \tag{3.1}$$

where the latter is the "noise" process for finite m. We shall study $N_m(t)$ as $m \to \infty$, and then approximate $Q(t)$ by

$$Q(t) \simeq mq(t) + \sqrt{m}\, N(t), \tag{3.2}$$

$N(t)$ being the limiting noise. Note that if m is made large our assumptions imply that demand rate, and hence queue size grow indefinitely. To obtain a sensible limit we must scale: $m\mu$ is now the departure or interaction completion, rate.

Argue as follows concerning changes in $Q(t)$ in a short time interval of length dt: $Q(t)$ moves deterministically by an amount $\{[m-Q(t)]\lambda(t)-m\mu\}dt$ except when $Q(t) = 0$; furthermore, the variance of the displacement is simply $\{[m-Q(t)]\lambda(t)+m\mu\}dt + o(dt)$. Assuming that we wish to approximate by a diffusion, write the stochastic differential equation

$$dQ(t) = \{[m-Q(t)]\lambda(t)-m\mu\}dt + \sqrt{\{[m-Q(t)]+m\mu\}}\, dW(t) \qquad (3.3)$$

where $dW(t)$ is the "differential" of a Wiener process (distributed like a normal random variable with mean zero and variance dt). Doob [3] addresses the mathematical problems that arise; also Cox and Miller [1], Gikhman and Skorokhod [5], and McKean [8].

Next apply the normalization (3.1). We have

$$dQ(t) = \sqrt{m}dN_m(t) + mdq(t) =$$

$$\{[m(1-q(t))-\sqrt{m}N_m(t)]\lambda(t)-m\mu\}dt$$

$$+ \sqrt{\{[m(1-q(t))-\sqrt{m}N_m(t)]\lambda(t)+m\mu\}}\, dW(t) \qquad (3.4)$$

After division by m we obtain

$$dN_m(t) + \sqrt{m}\, dq(t) = \{[\sqrt{m}(1-q(t)) - N_m(t)]\lambda(t) - \sqrt{m}\,\mu\}dt$$

$$+ \sqrt{\left[(1-q(t)) - \frac{N_m(t)}{m}\right]\lambda(t) + \mu}\, dW(t) \qquad (3.5)$$

Clearly if $m \to \infty$ the above equation makes sense only if coefficients of $\sqrt{m}$ cancel out; this leads to a differential equation for the deterministic part, $q(t)$:

$$\frac{dq}{dt} = \lambda(t)[1-q(t)] - \mu. \tag{3.6}$$

Taking $m \to \infty$ in the remaining equation yields the stochastic differential equation

$$dN(t) = -\lambda(t)N(t) + \sqrt{[(1-q(t))\lambda(t)+\mu]}\, dW(t). \tag{3.7}$$

The latter is recognized as describing a non-stationary Ornstein-Uhlenbeck diffusion process; see Cox and Miller [1], p. 225. We have neglected the boundary condition at zero (no queue can become negative) and hence our solution will be valid only when zero is seldom visited, i.e. when $\lambda(t) >> \mu$, at least most of the time. But this is precisely the peak load problem that is often of interest to designers and users.

Deterministic Component: Solution

The equation (3.6) may be solved by routine methods. We find

$$q(t) = q(0)e^{-\int_0^t \lambda(s)ds} + \int_0^t [\lambda(x)-\mu]e^{-\int_x^t \lambda(s)ds}\, dx \tag{3.8}$$

In case $\lambda(x) < \mu$ the relevant portion of the integrand should, in the spirit of our approximation, be replaced by zero.

Example: $\lambda(t) = \lambda$, a constant.

In this case, for $\lambda > \mu$

$$q(t) = q(0)e^{-\lambda t} + \left(1 - \frac{\mu}{\lambda}\right)(1-e^{-\lambda t}) \tag{3.9}$$

and in the long run $(t\to\infty)$

$$q(t) = 1 - \frac{\mu}{\lambda} \tag{3.10}$$

A little thought reveals the plausibility of this result. Our system is behaving like an infinite server system fed by a Poisson source, where the servers are now the terminals, and the computer is the source (of returned transactions). The expected number residing at the terminals is approximately $m\,\frac{\mu}{\lambda}$, the remainder waiting or being served at the computer.

Stochastic Component: Solution

A formal solution of (3.7) is seen to be

$$N(t) = N(0)e^{-\int_0^t \lambda(s)ds} + \int_0^t e^{-\int_x^t \lambda(s)ds}\,\sigma(x)dW(x) \tag{3.11}$$

where

$$\sigma(t) = \sqrt{[(1-q(t))\lambda(t)+\mu]} \tag{3.16}$$

and the integral involving $dW(\cdot)$ may be interpreted by integration by parts. Apparently $N(t)$ is normal or Gaussian with expectation $N(0)e^{-\int_0^t \lambda(s)ds}$ and, given $N(0)$, we find

$$E[(N(t)-N(0)e^{-\int_0^t \lambda(s)ds})^2] = \int_0^t e^{-2\int_x^t \lambda(s)ds}\,\sigma^2(x)dx \tag{3.13}$$

$$= \mathrm{Var}[N(t)]$$

Example: $\lambda(t) = \lambda$.

In this case (3.13) becomes

$$E[(N(t)-N(0)e^{-\lambda t})^2] = \int_0^t e^{-2\lambda(t-x)}\ \sigma^2(x)dx. \tag{3.14}$$

where

$$\sigma^2(t) = \mu(2-e^{-\lambda t}) + [1-q(0)]\lambda e^{-\lambda t} \tag{3.15}$$

thus

$$E[(N(t)-N(0)e^{-\lambda t})^2] = \frac{\mu}{\lambda}(1-e^{-2\lambda t}),$$

$$= \mathrm{Var}[N(t)]. \tag{3.16}$$

There is close agreement between the long-run $(t\to\infty)$ moments of $Q(t)$, and of the approximation (3.2) when $\lambda >> \mu$.

4. Applications of the Approximation.

The mathematical results obtained may be useful in assessing various aspects of system performance. We briefly describe these.

a) Distribution of $Q(t)$, the number of terminal customers waiting.

In our approximation, $Q(t)$ is normally distributed such that

$$E[Q(t)] \simeq mq(t), \tag{4.1}$$

with $q(t)$ given by (3.8), and $\mathrm{Var}[N(t)]$ given by substituting into (3.13). One should find such approximations useful for quick estimates of time-dependent behavior, and also for checks of simulation validity.

b) Distribution of total waiting time.

It may be of interest to describe the total accumulated waiting time of all applications for service during a fixed time period, e.g. one day:

$$\bar{Q}(t) = \int_0^t Q(r)dr$$

Our approximation amounts to writing

$$\bar{Q}(t) \simeq \int_0^t \{mq(r) + \sqrt{m}\, N(r)\}dr$$

which immediately implies that $\bar{Q}(t)$ is approximately normally distributed with parameters

$$E[\bar{Q}(t)] \simeq m \int_0^t q(r)dr$$

and, if $N(0) = 0$,

$$\mathrm{Var}[\bar{Q}(t)] \simeq m\, \mathrm{Var}\left[\int_0^t N(r)dr\right] \tag{4.2}$$

where the latter may be simplified by integration by parts.

Example: $\lambda(t) = \lambda$.

Evaluation of (4.2) requires an expression for the variance of the integral of $N(r)$. Integration by parts gives

$$\bar{N}(t) = \int_0^t N(r)dr = \int_0^t \left[\frac{1-e^{-\lambda(t-x)}}{\lambda}\right]\sigma(x)dW(x) \tag{4.4}$$

and thus

$$\mathrm{Var}[\bar{N}(t)] = \int_0^t \left[\frac{1-e^{-\lambda(t-x)}}{\lambda}\right]^2 \sigma^2(x)dx \tag{4.6}$$

If $q(0) = N(0) = 0$, then by (3.15)

$$\mathrm{Var}[\bar{N}(t)] = \int_0^t \left[\frac{1-e^{-\lambda(t-x)}}{\lambda}\right]^2 [2\mu+(\lambda-\mu)e^{-\lambda x}]dx$$

$$\sim \frac{2\mu}{\lambda^2}\, t \quad \text{as} \quad t \to \infty \tag{4.7}$$

Perhaps of more interest in the context of computer system applications, where interest focuses on t of the order of hours, and λ^{-1} on a one-minute scale, the above is expressible as

$$\mathrm{Var}[\bar{N}(t)] \sim \frac{2\mu}{\lambda^2}\, t \tag{4.8}$$

as $\lambda \to \infty$ for fixed t.

c) Probabilistic bounds on $Q(t)$.

Although we have found in a) an approximate normal distribution for $Q(t)$ at any <u>fixed</u> time t, it would be desirable to provide information on $Q(t)$ for <u>all</u> t during a particular period. For example,

we would like to be able to state that, during a day, queue size ever exceeds some specified level only with a given (small) probability. Such statements can be made if we can derive a boundary, $B(t)$, such that

$$P\{Q(t) \leq B(t), \quad \text{all} \quad t \leq T\}. \tag{4.9}$$

is specified.

In fact, such boundaries may be derived mathematically and we describe two of them. Unfortunately these are not entirely appealing from a practical viewpoint: either the boundary shape is unsatisfactory, or the probability associated with a boundary crossing is not expressed in a usable mathematical form.

c-1) Transformation to Wiener process with linear boundary.

Our first boundary problem is derived by expressing $N(t)$, the noise process, as a Wiener process, and referring the latter to a linear boundary. Probabilities for crossing such boundaries are well known.

Transform (3.7) into

$$d[N(t)e^{\int_0^t \lambda(s)ds}] = e^{\int_0^t \lambda(s)} \sqrt{[(1-q(t))\lambda(t)+\mu]}\, dW(t), \tag{4.10}$$

from which it is clear that the process $\{N(t)e^{\int_0^t \lambda(s)ds}\}$ is Wiener with zero mean and infinitesimal variance equal to $e^{2\int_0^t \lambda(s)ds}[(1-q(t))\lambda(t)+\mu]$. A time-scale transformation $\tau(t)$ converts the Wiener process into one with stationary increments; necessarily

$$d\tau = e^{2\int_0^t \lambda(s)ds}[(1-q(t))\lambda(t)+\mu]dt \tag{4.11}$$

or

$$\tau(t) = \int_0^t e^{2\int_0^r \lambda(s)ds} [(1-q(r))\lambda(r)+\mu]dr. \qquad (4.12)$$

Thus

$$N(t)e^{\int_0^t \lambda(s)ds} \equiv V(\tau), \qquad (4.13)$$

a zero-drift Wiener process with unit infinitesimal variance and $V(0) = N(0) = 0$. It is known (see Cox and Miller, pp. 220-221) that if $a > 0$

$$P\{V(\tau) \leq a + b\bar{\tau}, \ \forall\ \tau \leq \bar{\tau}\} = \Phi\left(\frac{a+b\bar{\tau}}{\sqrt{\bar{\tau}}}\right) - e^{-2ab}\,\Phi\left(\frac{-a+b\bar{\tau}}{\sqrt{\bar{\tau}}}\right), \qquad (4.14)$$

Φ being the standard normal distribution function.

Thus with the above probability, (4.11),

$$N(t) \leq \{a + b\int_0^t e^{2\int_0^r \lambda(s)ds} [(1-q))\lambda(r)+\mu]dr\}e^{-\int_0^t \lambda(s)ds}\ (t) \qquad (4.15)$$

for <u>all</u> $t, 0 \leq t \leq T$, and $\tau(t) = \bar{\tau}$. From this result we can state that with (approximate) probability (4.14)

$$Q(t) \leq mq(t) + \sqrt{m}\,\mathcal{B}(t). \qquad (4.16)$$

Example. $\lambda(t) \equiv \lambda$, a constant.

Let $q(0) = 0$. Then

$$\mathcal{B}(t) = \{a + b\int_0^t e^{2\lambda r}[2\mu+(\lambda-\mu)e^{-\lambda r}]dr\}e^{-\lambda t}$$

$$= ae^{-\lambda t} + b[(1 - \tfrac{\mu}{\lambda}) + \tfrac{\mu}{\lambda}\,e^{\lambda t} - e^{-\lambda t}] \qquad (4.17)$$

and hence

$$Q(t) \le m[(1 - \frac{\mu}{\lambda})(1-e^{-\lambda t})] + \sqrt{m}\, \mathcal{B}(t) \qquad (4.18)$$

for all $t \le T$;

$$\bar{\tau} = \int_0^T e^{2\lambda r}[2\mu+(\lambda-\mu)e^{-\lambda r}]dr = \frac{\mu}{\lambda}(e^{2\lambda T}-1) + (1 - \frac{\mu}{\lambda})(e^{\lambda T}-1) \qquad (4.19)$$

so for given T one solves for $\bar{\tau}$ required to evaluate (4.18). Unfortunately, the above boundary tends to be excessively U-shaped; in particular it is very high--and hence uninformative--for large T.

c-2) Straight line boundary for stationary Ornstein-Uhlenbeck noise.

Let us consider the special case of $\lambda(t) = \lambda$ a constant. Then $N(t)$, the limiting noise, eventually becomes Ornstein-Uhlenbeck, satisfying

$$dN(t) = -\lambda N(t)dt + \sqrt{2\mu}\, dW(t) \qquad (4.20)$$

or

$$dN(t') = -\lambda N(t')dt' + dW(t') \qquad (4.21)$$

if $t' = 2\mu t$. The time to reach a boundary $\mathcal{B}(t) = c > 0$ from x, denoted by T_c has a Laplace transform that may be expressed in terms of Weber, or parabolic cylinder, functions; according to Darling and Siegert [2]

$$E[e^{-sT_c}|N(0) = x] = \frac{D_{-s}(x)}{D_{-s}(c)}\ , \qquad (4.22)$$

but the actual probability distribution of T_c is not given numerically in the paper cited, and I have not seen a numerical tabulation in the literature. Actually, such a tabulation would be useful for one could then assess the probability distribution of $\max Q(t)$ over a one-day period.

5. Concluding Remarks.

The above model may be generalized in various directions, some of which are likely to be useful.

a) Service rate, μ, need not be a constant, but may be made a function of time. In particular, justification of the time-dependent results for arrival and service functions that are constant over t-intervals should be straightforward. A time-dependent service rate might be required to reflect the difference in business submitted (transaction type changes) throughout a day. Time dependent changes in m, the number of active terminals, may also be handled on a piecewise constant basis.

b) Service need not be exponential-Markovian. If traffic is heavy and transactions require arbitrarily independently distributed random times S to terminate, then it is natural to substitute the renewal theoretic asymptotic variance into the stochastic differential equation (3.7): simply replace μ by $\mathrm{Var}[S](E[S])^{-3}$ under the radical.

Finally, it is of interest that a variety of different, but similar, stochastic models for cyclic service result in the same limiting "noise" process to be superimposed upon the deterministic solution. This sort of robustness encourages one to think that simple models of the type suggested will often work adequately to generate useful insights, and to supplement or replace more complex calculations or simulations.

REFERENCES

[1] Cox, D. R. and Miller, H. D., The Theory of Stochastic Processes, John Wiley, New York, 1965.

[2] Darling, D. and Siegert, A. J. F., "The first passage problem for a continuous Markov process," Annals of Math. Stat, Vol. 24, 1953, pp. 624-639.

[3] Doob, J., Stochastic Processes, John Wiley, New York, 1952.

[4] Feller, W., An Introduction to Probability Theory and Its Applications, Vol. I, John Wiley, New York, 1957.

[5] Gikhman, I. I. and Skorokhod, A. V., Introduction to the Theory of Random Processes, W. B. Saunders Co., 1963.

[6] Gordon, W. J. and Newell, G. F., "Closed queueing systems with exponential servers," Ops. Research, Vol. 15, No. 2, March-April 1967, pp. 254-265.

[7] Iglehart, D., "Limiting diffusion approximations for the many server queue and the repairman problem." J. of Applied Probability, Vol. 2, 1965, pp. 429-441.

[8] McKean, H., Stochastic Integrals, Academic Press, New York, 1969.

[9] McNeil, D. and Schach, S., "Central limit analogues for Markov population processes," J. of the Royal Stat. Soc. (B), Vol. 35, No. 1, 1973.

INITIAL DISTRIBUTION LIST

	No. Copies
Defense Documentation Center (DDC) Cameron Station Alexandria, Virginia 22314	12
Library (Code 0212) Naval Postgraduate School Monterey, California 93940	2
Dean of Research (Code 023) Naval Postgraduate School Monterey, California 93940	1
Library (Code 55) Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	3
Professor D. R. Cox Department of Mathematics Imperial College London, England	1
Marvin Denicoff Office of Naval Research Arlington, Virginia 22217	1
Dr. Thomas Varley Office of Naval Research Arlington, Virginia 22217	1
Dr. D. Iglehart Operations Research Department Stanford University Stanford, California 94305	1
Dr. F. Hillier Operations Research Department Stanford University Stanford, California 94305	1
Dr. R. G. Miller Statistics Department Stanford University Stanford, California 94305	1

Dr. S. Stidham
Operations Research
Cornell University
Ithaca, New York 14850 — 1

Professor Leonard Kleinrock
Department of Computer Science
University of California
Los Angeles, California 90024 — 1

Dr. Mathew Sobel
Administrative Sciences
Yale University
New Haven, Connecticut 06520 — 1

Professor U. N. Bhat
Computer Science Center
Institute of Technology
Southern Methodist University
Dallas, Texas 75222 — 1

Dr. Daniel Heyman
Bell Telephone, Inc.
Crawford Corner Road
Holmdel, New Jersey 07733 — 1

Dr. Marcel Neuts
Statistics Department
Purdue University
West Lafayette, Indiana 47906 — 1

Dr. Ronald Wolff
Operations Research
University of California
Berkeley, California 94720 — 1

Dr. Gordon F. Newell, Transportation Engineering
University of California
Berkeley, California 94720 — 1

Dr. W. R. Schucany
Statistics Department
Southern Methodist University
Dallas, Texas 75222 — 1

Dr. H. L. Gray
Mathematics Department
Texas Tech University
Lobbock, Texas 79409 — 1

E. Ramras 1
Naval Personnel Research and Development Lab
San Diego, California 92100

Dr. Bruce McDonald 1
Office of Naval Research
Arlington, Virginia 22217

Professor Herbert Solomon 1
Statistics Department
George Washington University
Washington, D. C. 20006

Dr. D. R. McNeill 1
Statistics Department
Princeton University
Princeton, New Jersey 08540

Robert Agins 1
National Science Foundation
18th and "G" Streets
Washington, D. C.

Professor Gerald L. Thompson 1
GSIA
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Professor Jack McCredie 1
Computer Science
Carnegie-Mellon University
Pittsubrgh, Pennsylvania 15213

Professor John Lehoczky 1
Statistics Department
Carnegie-Millon University
Pittsburgh, Pennsylvania 15213

Professor Manuel Perlas 1
Department of Mathematics
Duquesne University
Pittsburgh, Pennsylvania 15200

Dr. P. A. W. Lewis 1
Dr. Kneale Marshall 1
Dr. R. Butterworth 1
Dr. Paul Milch 1
Dr. Donald P. Gaver 10
Department of Operations Research
and Administrative Sciences
Naval Postgraduate School
Monterey, California 93940

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and index ing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Postgraduate School Monterey, California 93940	Unclassified
	2b. GROUP

3. REPORT TITLE

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Donald P. Gaver

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1973	21	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

In this paper approximate models are given to describe backlogs of demands at a service system confronted by finitely many users. An application is to a time-shared computer center. Various functionals of the number waiting are discussed using the approximation

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Service systems						
Computers						
Time Sharing						
Stochastic models						